The effect of removing linguistic information upon identifying speakers of a foreign language

Schiller, N.O.; Koester, O.; Duckworth, M.

The effect of removing linguistic information upon identifying speakers of a foreign language

Niels O. Schiller,[*] *Olaf Köster*[†] *and Martin Duckworth*[**]

[*]*Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands;* [†]*University of Trier, Germany;* [**]*The College of St Mark & St John, Plymouth, England*

ABSTRACT The native-language background of a listener has been shown to have an effect upon identifying speakers of a foreign language. Previous experimental research showed that a German target speaker was identified significantly better when listeners were native speakers of German, or native speakers of English who had some knowledge of German, than when they did not know the language of the target speaker (Schiller and Köster 1996). This result was taken as support for the hypothesis that familiarity with the language of the target speaker has a positive effect upon identifying that speaker. This paper reports the results of a follow-up experiment that investigated the identification of a speaker in a voice line-up using nonsense speech. The results show that German natives, monolingual English natives, and English natives with some knowledge of German do not significantly differ in identifying a German speaker when most of the linguistic information of the language of the target speaker is removed from the stimulus materials.

KEYWORDS forensic phonetics, speaker identification, native-language background, voice line-ups, memory for voices.

INTRODUCTION

In forensic phonetics expert witnesses are more and more often involved in cases where they have to identify the voice of a foreign speaker (Huntley Bahr, Künzel, personal communications). However, until now very little is known about the effects of the native-language background on speaker identification (SID). The International Association for Forensic Phonetics (IAFP) advises phonetic expert witnesses to be cautious when working on speech samples in a language of which they are not native speakers. This statement implies that identifying voices in a foreign language is more difficult than identifying speakers of one's own native language. If that is the case, it might be legitimate to conclude that in SID listeners do not only rely on purely acoustic information – e.g., pitch, voice quality, and

1350–1771

Forensic Linguistics 4(1) 1997

spectral information – but also on linguistic information. (For an extensive discussion of potentially relevant parameters in SID see Hollien 1990; Künzel 1987; Nolan 1983). As Künzel (1994) points out, we do not yet know all the parameters that are potentially relevant in forensic SID. In this section, we review some of the literature suggesting that linguistic characteristics may play an important role in SID.

In an experimental study by Goldstein *et al.* (1981), recognition of unfamiliar voices was tested experimentally under different conditions. In particular, the listeners' task was to recognize speakers 'with and without accented voices' as well as speakers speaking in a foreign language.[1] In their first experiment, a group of speakers of General American English listened to a tape-recorded English sentence spoken by either a Chinese, a black American, or a white American male target speaker. The sentence had a length of fifteen words. In an immediate recognition test, participants then listened to an English test sentence uttered by four different speakers, one of whom was the target speaker. The proportions of correct identifications were not significantly different for the three groups of speakers, suggesting that accented voices are not more difficult to recognize than unaccented voices. However, there was a non-significant trend for the Chinese speakers to be identified worst.

In their second experiment, however, Goldstein *et al.* (1981) reduced the length of the speech material to which listeners were exposed before they performed the recognition task. The reduction of the speech stimulus to a single word had the effect of reducing the overall percentage of correctly recognized voices considerably, especially for the Chinese target voices. They were significantly less often identified correctly than the black and white American English voices. This result suggests that reduced length of stimulus materials has a greater effect on voice recognition for accented than for unaccented voices.

Goldstein *et al.* (1981) also tested listeners' recognition memory performance for voices speaking in a foreign language compared to English with a strong foreign accent. Speakers of General American English without any knowledge of Spanish listened to two native Spanish speakers saying either a sentence in accented English or in Spanish. After a retention interval of ten minutes, they listened to ten different voices saying a different token of the same sentence in the corresponding language condition (either accented English or Spanish). As a result, listeners identified Spanish-speaking voices equally as well as the same voices speaking accented English. This showed that the recognition of voices speaking a language not known to the listeners is no worse than the recognition of accented voices speaking the native language of the listeners. Furthermore, the results of the third experiment imply that knowing the language of a speech sample is not a prerequisite for speaker recognition. Goldstein *et al.* (1981: 20) conclude that 'voice recognition is just as good (or as poor) for foreign voices as it is for native voices'.

The study by Goldstein *et al.* (1981) supports the hypothesis that accented voices are more difficult to recognize than unaccented voices, but the evidence for a foreign-language effect in SID does not seem to be particularly strong. However, the conclusion that 'voice recognition is just as good (or as poor) for foreign voices as it is for native voices' seems premature to us since their second experiment revealed a significant effect for the Chinese speakers. Furthermore, their third experiment had an incomplete design since only 'foreign' voices were tested. Goldstein *et al.* (1981) did not include English-speaking voices in that experiment because, in their first experiment, they could not find any evidence that unaccented voices were recognized better than accented voices. Thompson (1987) did include this condition in his study. He had six male English–Spanish bilinguals producing three different versions of each of two different text passages. The first version was in English, the second in Spanish, and the third also in English, but this time produced with a heavy Spanish accent. English monolinguals without any knowledge of Spanish served as listeners. They were first exposed to the first passage of one speaker produced in one of the three language conditions. A week later, they listened to the second passage in the same language condition, but this time the passage was produced by six different speakers including the target speaker, i.e., the speaker whom they had heard producing the first passage a week earlier. The listeners' task was to identify the voice of the target speaker in the line-up. The line-up was presented three times. The results revealed a significant effect of language. With respect to the hit rate, English voices were recognized more often than English voices with a heavy Spanish accent. Accented voices were better recognized than Spanish voices. The false-alarm rates did not show any reliable differences between the three language conditions. Thompson (1987) stated that monolingual English-speaking participants identified voices speaking English more accurately than voices speaking Spanish, with recognition of accented voices being intermediate between the two.

The second experiment was designed as a replication of the first, but this time the target voice was absent from the line-up. There was no reliable difference between the three language conditions, neither with respect to the correct rejections nor to the false alarms. However, the false-alarm rate was notably high (0.56) under these circumstances.

In his third experiment, Thompson (1987) replicated the first experiment using only an English and a Spanish version of the text passages and reducing the retention interval between first and second presentation of the material to twenty-five minutes. The hit rate showed a reliable effect of language, i.e., voices speaking English were recognized much better by English monolinguals than voices speaking Spanish. The false-alarm rates, however, did not show any reliable differences between the conditions. According to Thompson (1987), this result shows that monolingual English speakers identified voices speaking English more accurately than (the same) voices speaking Spanish. He gave the following interpretation of the results. In his opinion, listeners identify voices of their own native language and their own dialect best because in this condition they are most able to identify speaker-specific idiosyncrasies. However, when exposed to a number of voices that show large deviations from the voices speaking their own dialect, then these deviations presumably override the subtle differences between the voices. The listener is no longer able to detect the idiosyncratic differences between the voices.

While the basic results of Thompson's (1987) study are straightforward, two points deserve criticism. First, Thompson analysed hit and false-alarm rates separately in each of his experiments. While this is a legitimate approach, a detection-theoretic analysis such as the one suggested in the Analysis section of this paper is more powerful because it takes hits and false alarms into account simultaneously. Second, in his third experiment, the overall false-alarm rate is 0.29 while the hit rate for voices speaking Spanish is 0.27. The false-alarm rate for voices speaking Spanish is not given however. It is clear that the sensitivity to discriminate targets from foils was poor for voices speaking Spanish. However, if hit and false-alarm rate were actually identical it would mean that discrimination sensitivity would be zero, i.e., listeners were not able to discriminate targets from foils when voices were speaking Spanish. This would be an interesting result for the research presented in this paper as well as in Schiller and Köster (1996) and Köster and Schiller (1997).

Thompson's results served as a starting point for further research on the role of language familiarity in voice identification. Goggin *et al.* (1991) carried out a series of identification experiments focusing on the listeners' familiarity with the language of the speakers. In their first experiment, six male English–German bilinguals produced an English and a German version of two different text passages. Speakers of General American English served as listeners. They first listened to one version of the first passage produced by a particular speaker (the target speaker). After a retention interval of five minutes, they were asked to identify the voice of the target speaker in a voice line-up. In this line-up, the voices of all six different speakers producing the second passage in the corresponding language condition were included. The line-up was presented three times. The results revealed a clear effect of language, i.e., there were significantly more correct identifications for the English than for the German samples. This means that monolingual English listeners identified voices speaking English better than the same voices speaking German.

In a second experiment, native German listeners without any knowledge of English listened to the utterances of the English–German bilingual speakers. In this case, voices were identified significantly better when they were speaking German than when they were speaking English. Goggin *et al.* (1991) interpreted this result as experimental support for the hypothesis that language familiarity plays an important role in identifying voices.

This hypothesis received further support from a third experiment. Six male English–Spanish bilinguals produced three versions of two different text passages, one in English, one in Spanish, and one in English with a strong Spanish accent. Two groups of listeners served as participants: a group of Spanish–English bilinguals and a group of English monolinguals. Participants listened to the first passage in a particular language condition, and after a retention interval of thirty minutes they performed an identification task. The task was to identify the voice of the speaker in a line-up of six different voices including the target speaker. In the line-up, participants listened to the second passage produced in the same language condition as the passage they heard first. The results of the line-up showed that monolingual English listeners correctly identified English voices significantly more often than foreign accented voices. The correct identification was lowest with Spanish voices. The bilingual listeners, however, showed no reliable difference with respect to correct identifications of voices in the three language conditions. This result showed that familiarity with the language of the speaker had a positive effect on identification.

In their 1991 paper, Goggin *et al.* computed the detection-theoretic sensitivity measure *d'*, but they did not carry out any statistical tests on the differences between the *d'* values in different experimental conditions. Analysis was carried out on the proportions of correct identifications (hit rates), which has the drawback that false alarms are not taken into account at the same time (see Analysis section below).

In an earlier experiment (Köster *et al.* 1995; Schiller and Köster 1996) we tested the hypothesis made by Goggin *et al.* (1991) that the native-language background of a listener plays an important role in SID. We showed that familiarity with the language of the speaker has an effect on the ability to identify a speaker. In a direct identification test, three different groups of listeners were asked to identify the voice of one speaker from a set of six different speakers (closed test). Listeners with a knowledge of German performed generally better than listeners without any knowledge of German (for details see Köster *et al.* 1995; and Schiller and Köster 1996). We concluded that speaker identification involves not only purely phonetic information but also linguistic information.

EXPERIMENT

This paper is about a control experiment that re-tested the results of the experiment reported in Schiller and Köster (1996). If it is the case that listeners use not only acoustic but also linguistic information when they are identifying a speaker, then listeners with different native-language backgrounds should perform equally well when linguistic information does not play a role. To test this hypothesis, we designed an experiment in which we tried to eliminate linguistic information as much as possible. In an additional experiment, Goggin *et al.* (1991) showed that distortion of the stimulus material severely affects speaker recognition. They used four different types of materials – i.e., normal text passages, mixed words, mixed syllables, and reversed speech – and found that the proportion of correct responses decreased as the passages became progressively more incomprehensible. Goggin *et al.* (1991) only tested English-dominant listeners. In other words, they did not test whether the foreign-language effect in SID disappears when the materials become incomprehensible.

Method

Participants

There was a total of seventy listeners divided into three groups. The first group consisted of thirty-one native German listeners. All of them were students at the University of Trier. In the second group, there were twenty-five English native speakers who had no knowledge of German. They were students of the College of St Mark & St John in Plymouth. The third group comprised fourteen native English listeners who had some knowledge of German. Participants from the third group took part in a university exchange programme at the University of Trier. All of them had studied German before they went to Germany. They were tested after having been in Germany for several months. All participants took part in the experiment voluntarily. None of them reported any hearing problems.

Materials

The speech materials used in this experiment came from six different German native speakers. They were recorded using a SONY ECM-737 unidirectional electret condenser microphone and a SONY TCD-D7 DAT recorder. Lip-to-microphone distance was approximately 30 cm. Speakers read a passage that was similar in length to the original passage used in the earlier experiment (140 words). But this time, all syllables of the original passage were substituted by the syllable /ma/. We assumed that this was an appropriate way to minimize the cues of the target language (German) in the materials. Some linguistic information, however, especially certain phonetic and phonological features such as the articulatory settings for German and prosodic features, were likely to have remained in the materials.

From each of the six speakers three parts of the passage were spliced out of the recordings using a wave form editor (Computerized Speech Lab, Kay Elemetrics Corporation), each between four and eight seconds in length. These speech samples were then recorded again through a telephone line so that we had six different samples from each speaker. Each of the six speech samples was re-recorded three times yielding a total of 108 samples (3 speech samples x 2 transmission conditions x 3 repetitions x 6 speakers = 108 stimuli). One speaker was chosen as the target speaker, the remaining five were foils. All speech samples were copied to a DAT tape in randomized order with the constraint that no two samples of the target voice occurred immediately adjacent.

Design

The three groups were tested separately. Listeners were first familiarized with the voice of the target speaker by listening five times to the whole text passage. Familiarization took approximately five minutes. Listeners were instructed to try to memorize the voice of the target speaker for recognition purposes. After the familiarization, there was a short break and response sheets were distributed to the participants. The retention period lasted approximately five minutes. Listeners were now given a forced-choice test. They listened to the DAT tape containing the 108 speech samples in a randomized order and were instructed to mark 'yes' on their response sheets whenever they recognized the voice of a speech sample as the one they had been familiarized with before. If they did not recognize the voice, they had to mark 'no'. For each trial, participants had five seconds to make their decision, then they heard the next trial. After every tenth speech sample, there was a sine tone of 300 Hz to help participants to keep track of the task. The entire voice line-up was presented only once and had a duration of approximately thirty minutes.

Analysis

Discrimination sensitivity

In a two-alternatives forced-choice (2AFC) test of the type used here, four different kinds of responses can be distinguished (Macmillan and Creelman 1991; McNicol 1972), namely hits, misses, false alarms, and correct rejections. Correctly recognizing the target voice is termed a hit, while failing to recognize it is called a miss. Recognizing a foil as the target by mistake is a false alarm. Correctly identifying a foil as such is a correct rejection (see Table 1).

Table 1 can be summarized by two numbers: the hit rate (H), i.e., the proportion of target trials to which the participant responded 'yes', and the false-alarm rate (F), i.e., the proportion of foil trials to which the participant incorrectly responded 'yes'. These two proportions can be used to determine the participants' sensitivity to the target–foil difference. By 'sensitivity' we mean the ability to discriminate between targets and foils. What participants do in a speaker-recognition experiment of the type presented in this study is to judge the familiarity of different voice samples. Under the assumption that none of the listeners was previously familiar with the voices of the speakers, samples of the target voice should be more familiar after exposure to that voice than are samples of any foil. Participants have to establish a criterion (k) that divides the familiarity dimension into two parts. Voice samples that fall above k are responded to with 'yes',

Table 1 Overview of the different response types in a 2AFC test

	response		
stimulus class	*'yes'*	*'no'*	*total*
target	hits	misses	number of targets
foil	false alarms	correct rejections	number of foils

below k voice samples are rejected ('no'). The criterion k is equally relevant for both the target and the foil voice samples. When sensitivity is high, target and foil voice samples differ greatly in average familiarity, and consequently the distributions of the target and foil voice samples in the decision space have very different means. When sensitivity is low, however, the means of the two distributions are closer together. The distance between the two means can therefore be understood as a measure of sensitivity.

The sensitivity measure provided by Signal Detection Theory (SDT) is called d' (Macmillan and Creelman 1991; Green and Swets 1966; McNicol 1972). It represents the distance between the means of the underlying distributions of the target and foil voice samples, in units of the common standard deviations. d' is a specific measure of the discrepancy between a hit rate and a false-alarm rate. d' is defined as the difference between the z-transformed hit and false-alarm rate:

$$d' = z(\mathrm{H}) - z(\mathrm{F})$$

The z-transformation is a standard procedure in statistics that is used to normalize proportions that come from different populations in order to make them comparable. Actually, the z-transformation translates H and F values into values of the zone of dispersion of the normal distribution. As a result, hit and false-alarm rates are converted to a z-score, i.e., to standard deviation units. A hit or false-alarm rate of 0.5 (50 per cent correct or incorrect) yields a z-score of 0, rates above 0.5 yield positive z-scores, and values below 0.5 are converted into negative z-scores. Two proportions that are equally far away from 0.5 yield the same absolute z-score.

If participants cannot discriminate between targets and foils, the hit rate equals the false-alarm rate, i.e., H = F. Accordingly, d' is 0 in this case. If the sensitivity is perfectly accurate, d' may be infinitely high, depending on the number of decimals to which H and F are carried. Usually, a d' value of 4.65 which results from a hit rate of 0.99 and a false-alarm rate of 0.01 is considered to represent a ceiling. The reason why d' is a good measure of sensitivity is that it depends both on H and on F, i.e., sensitivity increases when either H increases or F decreases. For the purpose of illustration, let us assume two participants (A and B) who took part in a discrimination experiment and yielded the following results: A achieved a hit rate of H = 0.8 and a false-alarm rate of F = 0.2, B yielded H = 0.8 and F = 0.3. If we look at both rates separately, it might seem as if both A and B performed equally well in the discrimination task because they achieved the same number of hits. However, B has a slightly higher false-alarm rate. A detection-theoretic analysis takes into account both H and F simultaneously. In our example, A would yield a d' value of 1.684, whereas for B d' would only be 1.366. To illustrate the fact that d' increases when either H increases or F decreases, let us assume that A had a hit rate of H = 0.9 while the false-alarm rate remains the same (F = 0.2). Under these circumstances, d' increases from 1.684 to 2.124. Similarly, d' yields the same value if the false-alarm rate decreases (F = 0.1) while the hit rate remains the same (H = 0.8; d' = 2.124).

Differences in discrimination performance are revealed by Receiver Operating Characteristic (ROC) curves, i.e., functions that relate a given hit and false-alarm rate. The major diagonal in the ROC space represents the *chance line*, i.e., H and F are equal. Points on a specific ROC curve have the same sensitivity, they only differ with respect to response bias. In SDT, the sensitivity measure d' represents the distance between a specific point on a ROC curve and the major diagonal.

To evaluate differences in sensitivity between two conditions, a 95 per cent confidence interval around the difference between the two d' values is determined (see Macmillan and Creelman 1991 for details). If zero is not in the interval, the two ROC points reflect significantly different sensitivities. For discrimination experiments involving data from many participants, Macmillan and Kaplan (1985) showed that detection-theoretic sensitivity measures can be computed from response rates averaged across subjects. In this study, collapsed d' values (d'_{coll}) were computed by averaging the hits and false alarms within each group of listeners. d'_{coll} was then calculated from the averaged proportions of hits and false alarms. A value of d' based on averaged data will generally be lower than the average d' of individual points because ROC curves are characterized by a continuously decreasing slope (see Macmillan and Kaplan 1985 for a more extensive discussion). However, the decrement severely affects averaged d' only if the two points are quite different with respect to response bias. Generally speaking, computing a collapsed d' from averaged data is a reliable way to estimate true average d' for group data. Individual participants' data that contain proportions of 0 or 1 are problematic for a detection-theoretic analysis because these proportions correspond to infinite d' values. In such cases, the experimenter has to decide on independent grounds whether these d' values are truly infinite or only statistically so. A statistically infinite d' may arise from a small number of trials that lead to no errors.

Response bias

Participants in a discrimination experiment may have a tendency to prefer one type of response over the other. Such a preference is called a *response bias*. It is generally assumed that the response bias is a monotonic function of both the hit and the false-alarm rate (Macmillan and Creelman 1991). In SDT, the basic bias measure is called c. c is computed by multiplying the sum of the z-transformed hit and false-alarm rate by the factor –0.5.

$$c = -0.5[z(\mathrm{H}) + z(\mathrm{F})]$$

This has the effect that c is negative when the false-alarm rate exceeds the miss rate, i.e., (1–H). This means that there is a *yes-bias*, i.e., a tendency to say 'yes'. Positive values for c arise when the false-alarm rate is lower than the miss rate. In this case, participants have a tendency to say 'no', i.e., there is a *no-bias*.

Results

Hits and false alarms were counted for each participant and then pooled across groups. There were a few cases of individual data that yielded perfect discrimination of the target voice. It is, however, sensible to assume that the corresponding d' values are only statistically but not truly infinite because the number of target voice trials was rather small (only eighteen, including three repetitions and two different transmission conditions). Results are presented for high fidelity and telephone transmission trials taken together (*All trials*), as well as for high fidelity and telephone transmission trials separately.

All trials

Group 1 (German natives) achieved 404 hits out of 558 target voice trials and 394 false alarms out of 2790 foil trials (see Table 2). This equals a hit rate of 0.72 and a false alarm rate of 0.14.

Table 2 Distribution of responses of group 1 (German natives, n = 31) for all trials

	response		
stimulus class	*'yes'*	*'no'*	*total*
target voice	404 (H = 0.72)	154	558
dummy voice	394 (F = 0.14)	2396	2790

Group 2 (monolingual English natives) made 335 hits out of 450 target voice trials and 501 false alarms out of 2250 foil trials (see Table 3). The corresponding hit rate is 0.79, the false-alarm rate is 0.22.

Table 3 Distribution of responses of group 2 (monolingual English natives, n = 25) for all trials

	response		
stimulus class	*'yes'*	*'no'*	*total*
target voice	355 (H = 0.79)	95	450
dummy voice	501 (F = 0.22)	1749	2250

Finally, group 3 (English natives with knowledge of German) made 213 hits out of 252 trials (H = 0.85) and 283 false alarms out of 1260 trials (F = 0.22) (see Table 4).

Table 4 Distribution of responses of group 3 (English natives with knowledge of German, n = 14) for all trials

	response		
stimulus class	*'yes'*	*'no'*	*total*
target voice	213 (H = 0.85)	39	252
dummy voice	283 (F = 0.22)	977	1260

The corresponding d' values are 1.633 for group 1, 1.578 for group 2, and 1.808 for group 3 (see Figure 1). None of the differences between the three groups are statistically significant.

The response bias c yielded values of 0.249 for group 1, –0.017 for group 2, and –0.132 for group 3. The differences in response bias were statistically significant between groups 1 and 2, as well as between groups 1 and 3, but not between groups 2 and 3 ($p < 0.05$).

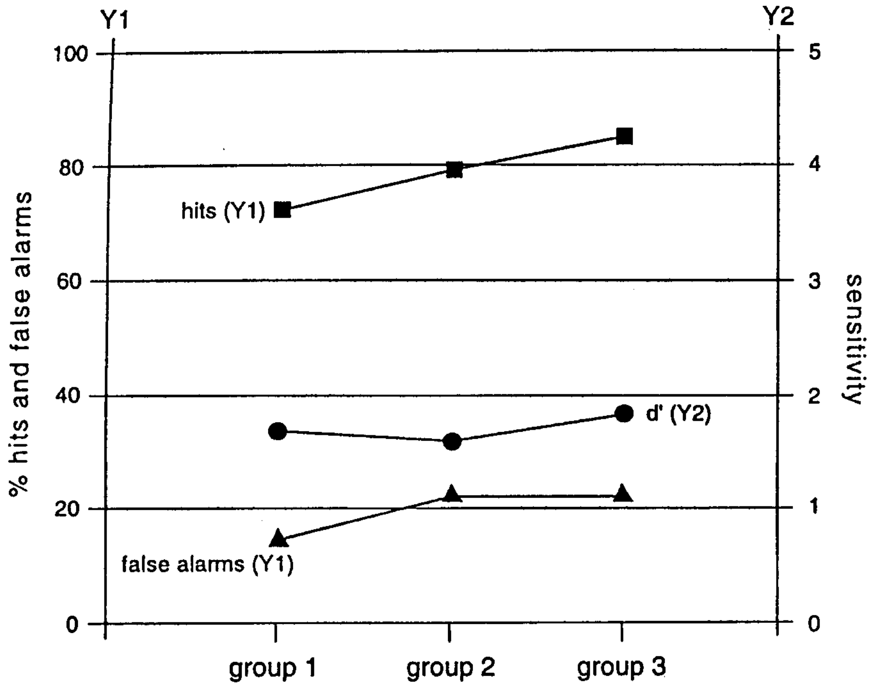

Figure 1 Hit rate (H), false-alarm rate (F), and sensitivity (*d'*) for the three listener groups

High-fidelity trials

Looking at the high-fidelity trials only, group 1 made 193 hits out of 279 high-fidelity target trials (H = 0.69) and 172 false alarms out of 1395 high-fidelity foil trials (F = 0.12) (see Table 5).

Table 5 Distribution of responses of group 1 (German natives, n = 31) for high-fidelity trials

	response		
stimulus class	*'yes'*	*'no'*	*total*
target voice	193 (H = 0.69)	86	279
dummy voice	172 (F = 0.12)	1223	1395

Group 2 made 182 hits (H = 0.81) and 301 false alarms (F = 0.27) (see Table 6).

Table 6 Distribution of responses of group 2 (monolingual English natives, n = 25) for high-fidelity trials

	response		
stimulus class	*'yes'*	*'no'*	*total*
target voice	182 (H = 0.81)	43	225
dummy voice	301 (F = 0.27)	824	1125

For group 3, we counted 110 hits (H = 0.87) and 132 false alarms (F = 0.21) (see Table 7).

The corresponding *d'* values were 1.671 for group 1, 1.491 for group 2, and group 3 yielded a *d'* of 1.932. The differences in sensitivity were not significant between groups 1 and 2, nor between groups 1 and 3 ($p < 0.05$), whereas the sensitivity between groups 2 and 3 was significantly different for the high-fidelity trials. Response bias *c* yielded the following values: 0.340 for group 1, –0.133 for group 2, and –0.160 for group 3. The differences in response bias reached statistical significance between groups 1 and 2, and also between groups 2 and 3, whereas the difference between groups 2 and 3 was not significant ($p < 0.05$).

Table 7 Distribution of responses of group 3 (English natives with knowledge of German, n = 14) for high-fidelity trials

	response		
stimulus class	*'yes'*	*'no'*	*total*
target voice	110 (H = 0.87)	16	126
dummy voice	132 (F = 0.21)	498	630

Telephone trials

The telephone trials can also be considered separately. For these trials, group 1 made 211 hits (H = 0.76) and 222 false alarms (F = 0.18) (see Table 8).

Table 8 Distribution of responses of group 1 (German natives, n = 31) for telephone trials

	response		
stimulus class	*'yes'*	*'no'*	*total*
target voice	211 (H = 0.76)	68	279
dummy voice	222 (F = 0.16)	1173	1395

Group 2 made 173 hits (H = 0.77) and 200 false alarms (F = 0.18) (see Table 9), and for group 3 there were 103 hits (H = 0.82) and 151 false alarms (F = 0.24) (see Table 10).

Table 9 Distribution of responses of group 2 (monolingual English natives, n = 25) for telephone trials

	response		
stimulus class	*'yes'*	*'no'*	*total*
target voice	173 (H = 0.77)	52	225
dummy voice	200 (F = 0.18)	925	1125

Table 10 Distribution of responses of group 3 (English natives with knowledge of German, n = 14) for telephone trials

	response		
stimulus class	*'yes'*	*'no'*	*total*
target voice	103 (H = 0.82)	23	126
dummy voice	151 (F = 0.24)	479	630

These hit and false-alarm rates yielded the following *d'* values: 1.700 for group 1, 1.654 for group 2, and 1.615 for group 3. Statistical analyses revealed no significant differences between any of the three groups for the telephone trials ($p < 0.05$). For the response bias *c*, we obtained the following value: 0.144 for group 1, 0.088 for group 2, and –0.105 for group 3. Statistically, the differences in response bias reached significance between groups 1 and 3 as well as between groups 2 and 3, but not between groups 1 and 2 ($p < 0.05$).

For each group, the difference between sensitivity for high fidelity and telephone trials was also analysed statistically, but it was not significant in any single case ($p < 0.05$).

Discussion

The main hypothesis tested in this study was confirmed. If linguistic information is largely removed from the stimulus materials, listeners with different native-language backgrounds perform similarly in identifying a target speaker from a set of six different speakers in a 2AFC test. More specifically, considering all trials taken together, the differences in sensitivity between the three groups of listeners did not reach statistical significance in any one case. The fact that the native English listeners without any knowledge of German (group 2) yielded the highest *d'* value and the fact that none of the three groups performed significantly better than any other group suggests that even the native Germans (group 1) did not profit from any prosodic information or the like that may have remained in the stimulus materials.

Looking at the trials of the two transmission conditions separately, the differences in sensitivity between groups were significant in only one case (group 2 vs. group 3 for the high-fidelity trials). Interestingly, for groups 1 and 2 the sensitivity was higher for the telephone trials, whereas for group 3 the reverse was the case. Statistical analyses, however, revealed that the differences between the sensitivity for high fidelity and telephone trials were not significant for any of the three groups. This may have different explanations. First, the fact that none of the groups performed significantly better for the high fidelity than for the telephone trials might be due to a ceiling effect. Although one has to be cautious when making comparisons between experiments, the difference in sensitivity between the native German group in this and in our earlier study (Schiller and Köster 1996) is enormous. Recognition ability seems to seriously decrease when linguistic information is removed from the speech samples. If we assume that performance for the high fidelity trials in this study was poor already, then it seems as if another deterioration of the material (e.g., telephone transmission condition) does not have an additive effect on the sensitivity. Second, it may be the case that the telephone transmission condition did not degrade the material seriously enough to yield a decrease in sensitivity in the first place. Most of the relevant acoustic information of /ma/ is unaffected by the low pass (3400 Hz) filtering process that occurs during telephone transmission. Unlike fricatives, /m/ does usually not have any high-frequency noise, and the vowel /a/ has both F_1 (750 Hz) and F_2 (1250 Hz) clearly below the filtering threshold.[2]

The results of the response-bias analysis are more difficult to interpret. A priori, we expected all three groups to show either no response bias at all or the same kind of response bias. However, the statistical analysis showed that group 1 had a no-bias that was significantly different from the yes-bias of groups 2 and 3. So far, we have not been able to find an explanation for this difference.

CONCLUSION

The results of the experiment reported in this paper can be interpreted as additional support for the hypothesis that language familiarity plays an important role in SID. Whereas previous research in this area investigated the effects of speech materials varying in language and of listeners with varying native-language backgrounds, the experiment discussed here investigated the effect of removing linguistic information from the speech materials. The fact that no reliable differences in discrimination sensitivity could be found between participants who knew the language of the target speaker and those who did not suggests that the foreign-language effect found earlier is real. Further research is needed in order to determine the influences of segmental and suprasegmental language-specific characteristics in SID.

ACKNOWLEDGEMENTS

The authors would like to thank Angelika Braun and an anonymous reviewer for many helpful comments on this paper. The research in this paper was supported by a grant from the International Association for Forensic Phonetics to Olaf Köster and Niels O. Schiller.

NOTES

[1] The term 'accented voice' in the terminology of Goldstein *et al.* (1981) refers to voices that have a foreign accent. Presumably, speakers who have a General American English accent were regarded as 'accentless'.

[2] The authors would like to thank Allen Hirson for bringing up this suggestion during the discussion of an oral version of this paper given at the 1996 Annual Meeting of the IAFP in Wiesbaden, Germany, 7–11 July 1996.

REFERENCES

Goggin, J. P., Thompson, C. P., Strube, G. and Simental, L. R. (1991) 'The role of language familiarity in voice identification', *Memory and Cognition*, 19, 448–58.

Goldstein, A. G., Knight, P., Bailis, K. and Conover, J. (1981) 'Recognition memory for accented and unaccented voices', *Bulletin of the Psychonomic Society*, 17, 217–20.

Green, D. M. and Swets, J. A. (1966) *Signal Detection Theory and Psychophysics*, New York: Wiley.

Hollien, H. (1990) *The Acoustics of Crime: The New Science of Forensic Phonetics*, New York: Plenum Press.

Köster, O., Schiller, N. O. and Künzel, H. J. (1995) 'The influence of native-language background on speaker recognition', in K. Elenius and P. Branderud (eds), *Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm, Sweden, 13-19 August, 1995 Vol. 3*, Stockholm: KTH and Stockholm University, 306–9.

Köster, O. and Schiller, N. O. (1997) 'Different influences of the native language of a listener on speaker recognition', present volume, 18–28.

Künzel, H. J. (1987) *Sprechererkennung: Grundzüge forensischer Sprachverarbeitung*, Heidelberg: Kriminalistik-Verlag.

Künzel, H. J. (1994) 'On the problem of speaker identification by victims and witnesses', *Forensic Linguistics*, 1, 45–57.

Macmillan, N. A. and Kaplan, H. L. (1985) 'Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates', *Psychological Bulletin*, 98, 185–99.

Macmillan, N. A. and Creelman, C. D. (1991) *Detection Theory: A User's Guide*, Cambridge: Cambridge University Press.

McNicol, D. (1972) *A Primer of Signal Detection Theory*, London: Allen & Unwin Ltd.

Nolan, F. (1983) *The Phonetic Bases of Speaker Recognition*, Cambridge: Cambridge University Press.

Schiller, N. O. and Köster, O. (1996) 'Evaluation of a foreign speaker in forensic phonetics: a report', *Forensic Linguistics*, 3(1), 176–85.

Thompson, C. P. (1987) 'A language effect in voice identification', *Applied Cognitive Psychology*, 1, 121–31.